DOCUMENT RESUME

ED 406 445 TM 026 424

AUTHOR Rodriguez, Maximo
TITLE Norming and Norm-Referenced Test Scores.
PUB DATE Jan 97
NOTE 25p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Collection; *Error of Measurement; Identification; *Norm Referenced Tests; Norms; Sample Size; *Sampling; *Scores; *Test Construction

ABSTRACT

Norm-referenced tests yield information regarding a student's performance in comparison to a norm or average of performance by similar students. Norms are statistics that describe the test performance of a well-defined population. The process of constructing norms, called norming, is explored briefly in this paper. Some of the most widely reported norm-referenced test scores are reviewed, and guidelines are provided for their interpretation. Nine steps for conducting a norming study, based on the work of Crocker and Algina (1986), are presented. These are: (1) identify the population of interest; (2) identify the most critical statistics that will be computed for the sample data; (3) decide on the tolerable amount of sampling error for one or more of the statistics in step 2; (4) devise a procedure for drawing a sample from the population of interest; (5) estimate the minimum sample size required to hold the sampling error within the specified limits; (6) draw the sample and collect the data; (7) compute the values of the group statistics of interest and their standard errors; (8) identify the types of normative scores that will be needed and prepare the normative conversion tables; and (9) prepare documentation of the norming procedure and guidelines for interpretation of the norms. Four categories of norm-referenced test scores (percentiles, standard scores, developmental scales, and ratios and quotients) are described. (Contains 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
MAXIMO RODRIGUEZ

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 406 445

NORMING AND NORM-REFERENCED TEST SCORES

Maximo Rodriguez

Texas A&M University 77843-4272

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January, 1997.

TM026424

BEST COPY AVAILABLE

Abstract

Norm-referenced tests yield information regarding a student's performance in comparison to a norm or average of performance by similar students. Norms are statistics that describe the test performance of a well-defined population. The process of constructing norms, called norming, is briefly explored in the present paper. Some of the most widely reported norm-referenced test scores are reviewed, and guidelines for their interpretation is provided.

Kubiszyn and Borich (1996) claimed that the purpose of testing is to provide objective data that can be used along with subjective impressions to make better educational decisions. They discussed two main types of tests used to make educational decisions: criterion-referenced tests and norm-referenced tests. Criterion-referenced tests provide information about a student's level of proficiency in or mastery of some skill or set of skills. This is accomplished by comparing a student's performance to a standard of mastery called a criterion. Such information tells us whether a student needs more or less work on some skills or subskills, but it says nothing about the student's performance relative to other students.

Norm-referenced tests, on the other hand, yield information regarding the student's performance in comparison to a norm or average of performance by similar students. Norms are statistics that describe the test performance of a defined group of pupils (Noll, Scannell & Craig, 1979). As Brown (1976) noted, potentially there are a number of possible norm groups for any test. Since a person's relative ranking may vary widely, depending upon the norm group used for comparison, Brown claimed that the composition of the norm group is a crucial factor in the interpretation of norm-referenced scores. Along similar lines, Crocker and Algina (1986, pp. 431-432) pointed out,

> The normative sample should be described in sufficient detail with respect to demographic characteristics (e.g., gender, race or ethnic background, community or geographic region, socioeconomic status, and educational background) to permit a test user to assess

whether it is meaningful to compare an examinee's performance to their norm's group.

The process of constructing norms is called norming. Mc Daniel (1994) argued that the result of norming a test is always a table that allows the user to convert any raw score to a derived score that instantly compares the individual with the normative group. Several types of norm-referenced scores (also called derived scores) have been discussed. Brown (1976) discussed four major types: percentiles, standard scores, developmental scales, and ratios and quotients. In the present paper issues related to norming are briefly examined. Additionally, some of the most commonly used norm-referenced scores are reviewed.

Norming

As stated earlier, norming is the process of constructing norms. Crocker and Algina (1986, p. 432) observed that the recommended procedures for conducting a norming study are similar regardless of whether the norms are for local or broader use. These authors suggested the following nine steps:

1.- Identify the population of interest (e.g., all students in a particular school district or all applicants for admission to a particular program of study or type of employment).

2.- Identify the most critical statistics that will be computed for the sample data (e.g., mean, standard deviation, percentile ranks).

3.- Decide on the tolerable amount of sampling error (discrepancy between the sample estimate and the population parameter) for one of more of the statistics

in step 2. (Frequently the sampling error of the mean is specified.)

4.- Devise a procedure for drawing a sample from the population of interest.

5.- Estimate the minimum sample size required to hold the sampling error within the specified limits.

6.- Draw the sample and collect the data. Document the reasons for any attrition which may occur. If substantial attrition occurs (e.g., failure of an entire school to participate after it has been selected into the sample), it may be necessary to replace this unit with another chosen by the same sampling procedure.

7.- Compute the values of the group statistics of interest and their standard errors.

8.- Identify the types of normative scores that will be needed and prepare the normative conversion tables.

9.- Prepare written documentation of the norming procedure and guidelines for interpretation of the normative scores.

Types of Sampling

Sampling techniques are usually classified into two broad categories: nonprobability sampling and probability sampling. Nonprobability sampling refers to samples of convenience (also termed accidental, accessible, haphazard, expedient, volunteer). Arguments in favor of nonprobability sampling typically are based upon feasibilty and economic considerations. In this type of sampling it is not possible to estimate sampling error. Thus, validity inferences to a population cannot be ascertained.

Conversely, probability sampling is one in which every individual in a specified

population has a known probability of selection, and random selection is used at some point or another in the sampling process. Crocker and Algina (1986) stated that norming a test on a nonprobability sample increases the likelihood of systematic bias in the examinees' performances. In contrast, the use of a probability sample in the norming study reduces the possibility of systematic bias in test scores, and makes it possible to estimate the amount of sampling error likely to affect various statistics calculated from these scores.

Types of Probability Sampling

Probability sampling generally comprises four types of sampling techniques: simple random sampling, systematic sampling, stratified sampling, and cluster sampling (see Cochran, 1977; Jaeger, 1984; Kish, 1965; Pehazur & Pedhazur-Schmelkin, 1991). As Pedhazur and Pedhazur-Schmelkin (1991, p. 321) noted,

> Although they differ in specifics of their sample designs, the various probability sampling methods are alike in that every element of the population of interest has a known nonzero probability of being selected into the sample, and random selection is used at some point or another in the sampling process.

Crocker and Algina (1986) likened simple random sampling to the process of assigning each member of the population of interest a unique number, writing each number on a separate piece of paper, putting all the slips of paper in a hat, and drawing from the hat a given number of slips. Each examineee whose number is selected is chosen for the sample. They pointed out however, that the process of selection is typically done

by choosing a random starting point in a random number table and selecting each examinee whose number appears sequentially in the list until the desired number of examinees for the sample is reached.

When one computes the mean or any other statistic for a norming sample, one obtains an estimate of that parameter in the population. This estimate is subject to sampling error. If all the possible samples of a given size were drawn from the population and the mean calculated for each sample, then it would possible to describe the sampling distribution of the mean. The standard deviation of this distribution of means is called the standard error of the mean (S_M). Fortunately, the S_M can be estimated on the basis of a single sample by the formula

$$S_M = (Sx^2 / n)^{1/2}$$

where

Sx^2 = variance of scores for the sample

n = sample size

As can be seen from this formula, the two determinants of the accuracy of the sample mean are the variance of the sample and the size of the group. Thus, the greater the variability, the larger the sample size needed to achieve a given level of sampling error.

Pedhazur and Pedhazur-Schmelkin (1991) argued that simple random sampling is not often used in research because of the many constraints associated with it. Difficulty to obtain lists and numbered list of elements of relatively large populations; population of interest residing in wide areas; and investigator interested in studying specific subgroups of the population are a few of such constraints.

Systematic sampling refers to a process of sampling in which, following a random starting point, every kth element is selected into the sample. Dividing the population size by the sample size yields k (K = N/n). A random number between 1 and k is selected for the starting point of the sampling. From there on, every kth element is chosen until the desired sample size is reached.

In stratified random sampling strategy the population of interest is first divided into nonoverlapping subdivisions, called strata, on the basis of one or more classification variables. Each stratum is initially treated independently. Thus, elements within each stratum are randomly selected and individual estimates (e.g., mean, proportion) are obtained. These estimates are then weighted to arrive at an estimate for the population parameters. According to Pedhazur and Pedhazur-Schmelkin (1991), the intent in stratified sampling is to reduce sampling variability by creating relatively homogeneous strata with respect to the dependent variable of interest. Therefore, as Crocker and Algina (1986) pointed out, stratified sampling allows the test developer to produce norms with less sampling error as would a simple random sample of comparable size.

Cluster sampling is used when sampling units are comprised of more than one element (e.g., classrooms, schools, factories, city blocks). These aggregates or clusters of elements are then randomly selected. In its simplest form, cluster sampling consists of sampling clusters only once and treating all elements of the selected clusters as comprising the sample. This is referred to as single-stage sampling. Conversely, in multistage sampling, selection proceeds in stages, each of which requires a different type of sampling frame from which appropriate clusters are drawn. For example, let us

suppose that a researcher is interested in conducting a norming study with a sample of fourth graders in a particular state. First, a random sample of counties is drawn. Second, within the counties selected, districts are randomly sampled. Third, within each district, schools are randomly drawn. Fourth, within the schools selected, fourth grade classrooms are randomly sampled. Finally, all fourth graders within the classrooms selected comprise the sample. Alternatively, fourth graders may be randomly selected within classrooms.

Describing the Norming Study in the Test Manual

Crocker and Algina (1986) claimed that the test developer must include several crucial pieces of information in the description of a norming study. First, a description of the population for whom the test is intended. Second, a complete documentation of the procedure by which the norming sample was selected (i.e., sampling plan, including a description of the type of sampling technique used, refusal and/or nonresponse rate). Third, one must report the date of the norming study with a detailed description of the norming group in terms of gender, racial or ethnic background, socioeconomic status, geographic location, and types of communities represented. Fourth, statistics computed to describe the performance of the norming group on the test (e.g, mean, proportion, standard deviation), accompanied by information of their accuracy--at least, the standard error of the mean-- should be reported. Finally, clear explanations of the meanings and appropriate interpretations of each type of normative score conversion should be reported.

Norm-referenced Test Scores

As said earlier, norming studies are typically conducted to construct conversion tables so that an individual's raw score can be compared to the score of other individuals in a relevant reference group, the norm group. In the following sections some of the most common types of norm-referenced or derived scores will be described. Although there are a number of possibly ways of classifying derived scores (see, e.g., Angoff, 1971; Lyman, 1971, Nunally, 1964), Brown's four-way classification--percentiles, standard scores, developmental scales, and ratios and quotients--will be adopted.

Percentiles

Percentiles are among the most widely used derived scores because of their ease of interpretation. Although some authors use the term "percentile" and "percentile rank" interchangeably, Mehrens and Lehman (1984, p. 318) distinguished between the two:

> A percentile is defined as a point in the distribution below which a certain percentage of the scores fall. A percentile rank gives a person's relative position or the percentage of students' scores falling below his obtained score. For example, the 98th percentile is the point below which 98 percent of the scores in the distribution fall. This does not mean that the student who scored at 98th percentile answered 98 percent of the items correctly.

Hinkle, Wiersma, and Jurs (1994, p. 52) also distinguished between percentile and percentile rank:

> Percentile rank of a score is the percentage of scores less than or equal to that score. For example, the percentile rank of 63 is the

> percentage of scores in the distribution that falls at or below a score of 63. It [percentile rank] is a point in the percentile scale, whereas a percentile is a score, a point on the original measurement scale. Mathematically, the percentile rank is defined as

$$P = [\, cfi + .5\,(fi) / N \,] \times 100\ \%$$

where

cfi is the cumulative frequency for all scores lower than the score of interest,

fi is the frequency of scores in the interval of interest,

N is the number in the sample.

Crocker and Algina (1986, pp. 439-440) described the basic steps in computing percentile ranks for a raw score distribution as follows:

1.- Construct a frequency distribution for the raw scores.

2.- For a given raw score, determine the cumulative frequency for all scores lower than the score of interest.

3.- Add half the frequency for the score of interest to the cumulative frequency value determined in step 2.

4.- Divide the total by N, the number of examinees in the norm group and multiply 100%.

Hinkle, Wiersma, and Jurs (1994) offered general formulas for computing either percentiles or percentile ranks when raw scores are grouped into class intervals. The formula for calculating percentiles is the following:

$$Px = ll + [(\, np - cf) / fi \,]\, w$$

where

ll = exact lower limit of the interval containing the percentile point

n = total number of scores

p = proportion corresponding to the desired percentile

cf = cumulative frequency of scores below the interval containing the percentile point

fi = frequency of scores in the interval containing the percentile point

w = width of class interval

The formula for computing percentile ranks is as follows

$$P_R = \{ [cf + (x - ll / w) fi] / n \} 100$$

where

x = score for which the percentile rank is to be determined

cf = cumulative frequency of scores below the interval containing the score x

ll = exact lower limit of the interval containing x

w = width of class interval

fi = frequency of scores in the interval containing x

n = total number of scores

Despite their ease of interpretation, percentile ranks have some major limitations that merit the attention of test users (Thompson, 1993). Brown (1976) discussed two of such limitations. First, being on an ordinal scale, percentile ranks cannot legitimately be added, subtracted, multiplied, or divided. According to this author, this is not a serious limitation when interpreting scores, but it is a serious liability in statistical analyses. A second limitation is, in his view, of more concern to the test user. Percentile ranks have a

rectangular distribution, whereas test score distributions generally approximate the normal curve. As a consequence, small raw score differences near the center of the distribution result in large percentile difference. Conversely, large raw score differences at the extremes of the distribution produce only small percentile differences.

Brown warned us that "unless these relations are kept in mind, percentile ranks can easily be misinterpreted, in particular, seemingly large differences in percentile ranks near the center of the distribution tend to be overinterpreted" (1976, p. 184).

Crocker and Algina (1986, p. 441) noted that the nonlinear conversion implicit in conversion to percentile ranks can cause people to misinterpret these scores:

> Most misinterpretations arise when test users fail to recognize that the percentile rank scale is a nonlinear transformation of the raw score scale. Simply put, this means that at different regions on the raw score scale, a gain of 1 point may correspond to gains of different magnitudes on the percentile rank scale.

Standard Scores

Brown (1976) argued that when statistical analyses are performed on test scores, it is desirable to have scores expressed on an interval scale--a scale with equal-size units. Standard scores have this property. Hopkins and Stanley (1981, p. 52) defined standard scores as "scores expressed in terms of a standard, constant mean and a standard, constant standard deviation." Standard scores are obtained by dividing each deviation score (subtracting the mean raw score from each raw score) by the standard deviation of the particular distribution:

$$z = x - X / s$$

where

z = the standard score

x = the raw score

X is the mean raw score

s is the standard deviation of the distribution.

Properties of Standard Scores

Brown (1976, p. 185) discussed the following five properties of standard scores:

1.- They are expressed as a scale having a mean of 0 and a standard deviation of 1.

2.- The absolute value of a z score indicates the distance of the raw score from the mean of the distribution. The sign of the z scores indicate whether the raw score falls above or below the mean; scores above the mean will have positive signs; scores below the mean, negative signs.

3.- Inasmuch as standard scores are expressed on an interval scale, they can be subjected to algebraic operations.

4.- The transformation of raw scores to standard scores is linear. Thus, the shape of the distribution of z scores is identical to the distribution of raw scores.

5.- If the distribution of raw scores is normal, the range of z scores will be from approximately -3 to +3.

Brown argued that if the distribution of standard scores is normal, standard scores can be directly converted into percentile ranks. This transformation can be made using a table of areas of the normal curve. This transformation is possible because in a normal

distribution there is a specifiable relationship between standard scores (z scores) and the areas within the curve (i.e., the proportion of cases falling between any two points).

Additionally, this author argued that even when raw scores are not normally distributed, it is possible to make an area transformation, and force scores into a normal distribution. Scores derived in this manner are called normalized scores; the word "normalized" indicates that scores have been forced into a normal distribution. In his view, to normalize scores, there must be some basis for assuming that scores on the characteristic being measured are, in fact, normally distributed. If scores cannot be assumed to be normally distributed, forcing them into normal distribution only distorts the distribution. Therefore, according to Brown, normalized standard scores should be computed only when an obtained distribution approaches normality, but because of sampling errors, is slightly different.

Whether standard or normalized, z scores have the disadvantage of assuming decimal and negative values, which can be difficult to interpret, particularly to people who are not familiar with educational measurement. As Nunally (1964, p. 46) observed,

> Although standard scores are directly useful to anyone who is familiar with educational measurement, people who are naive in this respect have some difficulty in interpreting standard scores. For example, a standard score of zero is often misinterpreted as meaning zero instead of average performance on the test. Some people find it difficult to understand negative standard scores, those below the mean. For these reasons, standard scores often are

transformed to a distribution having a desired mean and standard deviation.

Transformed Z Scores

Thus to avoid decimals and negative values, z scores are transformed to another scale. This transformation is of the form:

$$Y = m + k\,(z)$$

where

Y = the derived score

m and k = constant values arbitrarily chosen to suit the convenience of the test developer.

The constant m will transform the mean, and k the standard deviation. This linear transformation does not change the shape of the z score distribution. Transformed z scores include T scores, College Entrance Examination Board (CEEB) scores, Normal Curve Equivalent (NCE) scores, Deviation IQ scores, and Stanines.

A T score is a standard score with a mean of 50 and a standard deviation of 10. Thus, the general formula for the T score is

$$T = 50 + 10\,(z)$$

Since scores are not likely to fall more than 5 standard deviations below the mean, negative scores are eliminated. Additionally, multiplying the standard deviation by 10 eliminates decimals. Thus, a z score of -2 would convert to a T score of 30 and a z score of 1.7 would convert to a T score of 67.

The CEEB score scale, developed by the Educational Testing Service, has a mean of 500 and a standard deviation of 100. This score scale takes the form

$$Y = 500 + 100\,(z)$$

The convertion of the CEEB scale to either T score or z score is straightforward. For example, a score of 700 on the CEEB scale is equivalent to a T score of 70 and to a z score of +2. Each of these three standard scores indicates that the individual' score is 2 standard deviations above the mean.

Based on the general formula for deriving CEEB scores, a CEEB score of 500 under normal circumstances would indicate that the individual's score is right at the mean. However, as McDaniel (1994, pp. 100-101) pointed out, "We know that as of the fall of 1993, the Educational Testing Service reported that the average score for college-bound seniors on the verbal test was 424 and the average score for the mathematics test was 478." McDaniel explained this contradiction by arguing that the CEEB standard score scale was established in 1941 on the basis of the average performance taking the test at that time. Those students were primarily young men and women applying to prestigious and highly selective colleges, which required the test as part of the admission requirement. Now many colleges require the test and a much broader segment of the population is taking the test. This is, in his opinion, almost a classic case of a shift in the norm group. McDaniel claimed that although the standard scores for the Scholastic Aptitude Tests are still reported on the 1941 scale, the percentile scores based on students tested during the current year is a much better indication of performance on the tests.

Normal Curve Equivalent (NCE) scores are being reported by a number of test publishers. NCE scores are derived by converting percentile ranks to normalized z score and making a transformation of the form

$$NCE = 50 + 21.06$$

Thus, the NCE scale has a mean of 50 and a standard deviation of 21.06. According to McDaniel (1994) this rather strange standard deviation was chosen because it leads to NCE scores in which one corresponds to a percentile rank of 1 and ninety-nine corresponds to a percentile rank of 99. However, this author showed that anchoring the NCE scores to percentile ranks at these two points may not have been worth the effort since the two scores cannot be interpreted in the same way. NCE scores are on an interval scale, and in contrast to percentile ranks, NCE scores are meaningfully subjected to arithmetic operations such as calculating averages, making comparisons, and so forth.

The stanine is a nine-unit standard scale with a mean of 5 and a standard deviation of 2. Each unit, except units 1 and 9, is .5 standard deviation in width. This standard scale was developed by the United States Army Air Forces and used extensively during the World War II. Hopkins and Stanley (1981) suggested a set of procedures for converting raw scores to stanines:

1) Rank raw scores from the highest to the lowest.
2) Assign the top 4 % a stanine of 9.
3) The next 7 % are assigned a stanine of 8.
4) The next 12 % are assigned a stanine of 7.
5) Assign the next 17 % a stanine of 6.
6) The next 20 % are assigned a stanine of 5.
7) Use the same procedure to assign stanines 1, 2, 3, and 4 respectively.

Bauman (1988), in his discussion of the stanine scale, claimed that stanines have the advantage of being easily interpretable since each is a single digit; of being directly comparable across tests; and of being evenly spread out with respect to raw scores. However, he readily pointed out that stanines are rather gross measures. He argued that, for example, the exact percentile score for a student who obtained a 5th stanine on a test could range from 40 to 60, a rather large range.

Deviation Intelligence Quotient (DIQ) score is perhaps the most well-known of all transformed z scores. This scale replaced the IQ ratio (e.g., McDaniel, 1994; Mehrens & Lehmann, 1984). Typically, deviation IQs have a mean of 100 and a standard deviation of 15 or 16. However, Mehrens and Lehmann (1984) pointed out that standard deviations vary from test to test, ranging from as low as 12 to as high as 20. This is one of the reasons why these authors suggested that two individuals' IQ scores be compared only if they have taken the same test.

Developmental Scales

Developmental scales compare an individual's performance to that of the average person of various developmental levels. Typically, these scales report performance as grade or age equivalent. Grade Equivalent (GE) scores provide information about how a child's performance compares to that of other children at various grade levels. A GE score consists of one or two digits followed by a decimal point and another digit, such as 3.9, 7.0, or 10.2. The first digit represents the year in school; the digit following the decimal point represents the month in school. Thus, if a third-grader obtained a GE of 3.9 on a reading comprehension subtest, the score means that the student performed as well on that test as did the average student in the ninth month of third grade.

Mehrens and Lehmann (1984, pp. 322-323) discussed four major limitations of GEs scores. The first limitation is the problem of extrapolation. If for example, a particular sample is used in grades 4, 5, and 6, the curve showing the relationship between raw scores and GEs can be extrapolated so that the median raw scores for the other grade levels would be guessed. Mehrens and Lehmann claimed that the extrapolation procedure is based on the very unrealistic assumption that there would be no points of inflection (that is, no change in direction) in the curve if real data were available. An additional problem of extrapolation relates to sampling error. In these authors' view, small sampling errors can make extrapolated GEs very misleading.

A second limitation of GEs is that they give little information about the percentile standing of the person within the class. A fifth grader may, for example, because of the difference in the grade equivalent distributions for various subject matters, have a GE of 6.2 in English and 5.8 in mathematics and yet have a higher percentile rank in

mathematics. The third limitation of GEs is that (contrary to what the numbers indicate) a fourth-grader with a GE of 7.0 does not necessarily know the same amount or the same kinds of things as a ninth-grader with a GE of 7.0. The fourth limitation of GEs is that they are a type of norm-referenced measure particularly prone to misinterpretation by critics of education. Norms are not standards, and even the irrational critics of education do not suggest that everyone should be above the 50th percentile. Yet people talk continually as if all sixth-graders should be reading at or above the sixth-grade equivalent (for similar views, see Bauman, 1988; Crocker & Algina, 1986).

Age Equivalent (AE) scores are analogous to GE scores. The difference is that AE scores compare an individual's performance with that of persons of different ages, whereas GE scores compare an individual's performance with average student performance in various grades.

Ratios and Quotients

There have been numerous attempts to develop scales that use the ratio of two scores. The most popular score ratio is the intelligence quotient (IQ). The IQ, defined as the ratio of the child's mental age to his chronological age, was proposed as an index of the rate of intellectual development:

$$IQ = (MA / CA) \times 100$$

where

MA = mental age

CA = chronological age

As can be seen from this formula, a child whose mental age and chronological age are equal will obtain an IQ of 100, and will be judged to have an average intellectual development for this age. Similarly, a child whose mental development is more rapid than average will obtain an IQ over 100, whereas a child whose mental development is slower than average will obtain an IQ below 100. As Brown (1976, p. 194) noted,

> Because of nonequivalent standard deviations, and the fact that intellectual growth does not increase linearly with increasing age, ratio IQs are no longer used on major intelligence tests. Instead, normalized standard scores based on a representative sample of the population at each level are now used. These scores called deviation IQs, have a mean of 100 and a standard deviation of 15 (Weschler scales) or 16 (Stanford-Binet) points at each age level.

Mehrens and Lehmann (1984, p. 324) discussed two major weaknesses of IQs: First, the standard deviations of the IQS are not constant for different ages, so that an IQ score of say 112 would be equal to a different percentile at one age than at another. Second, opinions varied about what the maximum value of the denominator should be. When does a person stop growing intellectually--at 12 years, 16 years, 18 years? Because of these various inadequacies of the ratio IQ, these authors argued, most test constructors now report deviation IQs.

Another quotient score reported in a number of norms is the Educational Quotient (EQ). This ratio is intended to indicate the rate of educational development or achievement. EQ is obtained by dividing educational age (EA) by chronological age (CA) and multiplying the result by 100. Brown (1976) argued that educational or achievement ratios have two major drawbacks. First, the ratio of two unreliable scores will be less reliable than either individual measure. Thus, the quotient will, typically be a statistically unsound measure. Second, comparing a measure of achievement to one of intellectual ability assumes that achievement is determined solely by intellectual ability. In his opinion, this assumption is both constricting and inconsistent with empirical facts.

Summary

In the present paper a brief discussion of norms and of the process of norming was presented. It was argued that norm-referenced test scores are useful when test users are interested in comparing a student's score to a norm or average of performance by similar students. Nine steps were suggested to conduct a norming study. Additionally, it was argued that probability sampling allows the test developer to estimate the degree of

sampling error and reduces the likelihood of systematic bias in the normative data. Four different types of probability sampling were discussed. Finally, four categories of norm-reeferenced test scores: percentiles, standard scores, developmental scales, and ratios and quotients, were described.

REFERENCES

Angoff, W. (1971). Norms, scales, and equivalent scores. In R Thorndike (Ed.). Educational measurement (2nd ed.). Washington, D.C.: American Council on Education.

Bauman, J. (1988). Reading assessment: An instructional decision-making perspective. New York: Macmillan Publishing Company.

Brown, F. (1976). Principles of educational and psychological testing (2nd ed.). New York: Holt, Rinehart and Winston.

Cochran, W. (1977). Sampling techniques (3rd ed.). New York: Wiley.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Hinkle, D., Wiersma, W., & Jurs, S. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston: Houghton Mifflin Company.

Hopkins, K., & Stanley, J. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood, NJ.: Prentice Hall.

Jaeger, R. (1984). Sampling in education and the social sciences. New York: Longman.

Kish, L. (1965). Survey sampling. New York: Wiley

Kubiszyn, T., & Borich, G. (1996). Educational testing and measurement (5th ed.). New York: Harper Collins College Publishers.

Lyman, H. (1971). Test scores and what they mean (2nd ed.). Englewood Cliffs, NJ.: Prentice Hall.

McDaniel, E. (1994). Understanding educational measurement. Madison, Wisconsin: Brown & Benchmark Publishers.

Mehrens, W., & Lehmann, I (1984). Measurement and evaluation (3rd ed.). New York: CBS College Publishing.

Noll, V., Scannell, D., & Craig, R. (1979). Introduction to educational measurement (4th ed.). Boston: Houghton Mifflin Company.

Nunally, J. (1964). Educational measurement and evaluation. New York: McGraw Hill Book Company.

Pedhazur, E., & Pedhazur-Schmelkin, L. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ.: Lawrence Erlbaum Associates, Publishers.

Thompson, B. (1993, November). GRE percentile ranks cannot be added or averaged: A position paper exploring the scaling characteristics of percentile ranks, and the ethical and legal culpabilities created by adding percentile ranks in making "high-stakes" testing decisions. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Service N^{0}. ED 363 637)